BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

AUTOMATIC DOCUMENT CLASSIFICATION DEVICE, LEARNING DEVICE, CLASSIFICATION DEVICE, AUTOMATIC DOCUMENT CLASSIFICATION METHOD, LEARNING METHOD, CLASSIFICATION METHOD AND STORAGE MEDIUM

Patent Number: JP11085796
Publication date: 1999-03-30
Inventor(s): OTANI NORIKO; ITO SHIRO; SHIBATA SHOGO; UEDA TAKANARI; IKEDA YUJI
Applicant(s):: CANON INC
Requested Patent: JP11085796
Application JP19970250125 19970901
Priority Number(s):
IPC Classification: G06F17/30 ; G06F17/21
EC Classification:
Equivalents:

Abstract

PROBLEM TO BE SOLVED: To provide an automatic document classification device which can appropriately classify a document where other topics different from a subject appear.
SOLUTION: The automatic document classification device refers to a valid word dictionary and obtains paragraph vectors on a learning document and a document being a classification object (paragraph vector calculation part 105). An other topic paragraph is decided from the distribution of the paragraph vectors (other topic paragraph decision part 107) and the valid paragraph vector is taken out from the paragraph vectors by referring to the other topic paragraph. A document vector is obtained from the paragraph vector (document vector calculation part 109). In a learning phase, the folder vectors of respective categories are obtained by using the document vector of the learning document (folder vector calculation part 111). In a classification phase, the category to which the document of the classification document belongs is decided in accordance with a comparison result between the document vector of the document being the classification object and the folder vectors of the respective categories (classification decision part 113).

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－85796

(43)公開日　平成11年(1999)３月30日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401		３１０Ｄ
17/21			15/20	５９０Ｅ

審査請求　未請求　請求項の数９　ＦＤ　（全 14 頁）

(21)出願番号　特願平9－250125

(22)出願日　平成９年(1997)９月１日

(71)出願人　000001007
キヤノン株式会社
東京都大田区下丸子３丁目30番２号

(72)発明者　大谷　紀子
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内

(72)発明者　伊藤　史朗
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内

(72)発明者　柴田　昇吾
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内

(74)代理人　弁理士　渡部　敏彦

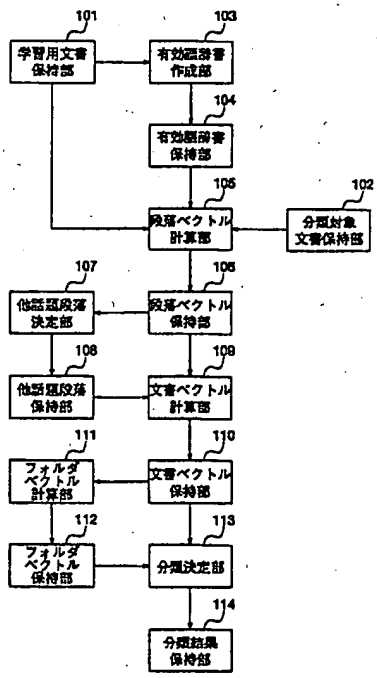
最終頁に続く

(54)【発明の名称】　文書自動分類装置、学習装置、分類装置、文書自動分類方法、学習方法、分類方法および記憶媒体

(57)【要約】

【課題】　主題と異なる他話題が出現する文書に対してその分類を適正に行うことができる文書自動分類装置を提供する。

【解決手段】　文書自動分類装置は、学習用文書と分類対象文書とのそれぞれについて、有効語辞書を参照して段落ベクトルを求め（段落ベクトル計算部１０５）、その段落ベクトルの分布から他話題段落を決定し（他話題段落決定部１０７）、その他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、この段落ベクトルから文書ベクトルを求める（文書ベクトル計算部１０９）。学習フェーズでは、学習用文書の文書ベクトルを用いて各カテゴリのフォルダベクトルを求め（フォルダベクトル計算部１１１）、分類フェーズでは、分類対象文書の文書ベクトルと各カテゴリのフォルダベクトルとの比較結果に応じて分類対象文書が属するカテゴリを決定する（分類決定部１１３）。



【特許請求の範囲】

【請求項1】　学習用文書と該学習用文書から選出された有効語を集めて作成した有効語辞書とを用いて、分類対象文書をユーザの意図に沿って分類する文書自動分類装置において、前記学習用文書と前記分類対象文書とのそれぞれについて、前記有効語辞書を参照して段落ベクトルを求める段落ベクトル計算手段と、前記学習用文書と前記分類対象文書とのそれぞれについて、その段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する他話題段落決定手段と、前記学習用文書と前記分類対象文書とのそれぞれについて、その他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める文書ベクトル計算手段と、前記学習用文書について求められた文書ベクトルを用いて各カテゴリのフォルダベクトルを求めるフォルダベクトル計算手段と、前記分類対象文書について求められた文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する分類決定手段とを備えることを特徴とする文書自動分類装置。

【請求項2】　分類対象文書をユーザの意図に沿って分類する文書自動分類システムに用いられる、前記分類対象文書が属するカテゴリを決定するための基準を求めるための学習装置において、学習用文書を保持する学習用文書保持手段と、前記学習用文書から有効語を選出し、該選出された有効語を集めて有効語辞書を作成する有効語辞書作成手段と、前記学習用文書について前記有効語辞書を参照して段落ベクトルを求める段落ベクトル計算手段と、前記学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する他話題段落決定手段と、前記学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める文書ベクトル計算手段と、前記学習用文書の文書ベクトルを用いて前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるフォルダベクトル計算手段とを備えることを特徴とする学習装置。

【請求項3】　分類対象文書をユーザの意図に沿って分類する文書自動分類システムに請求項2記載の学習装置とともに用いられる、前記分類対象文書が属するカテゴリを決定するための分類装置において、前記分類対象文書を保持する分類対象文書保持手段と、前記分類対象文書について前記有効語辞書を参照して段落ベクトルを求める段落ベクトル計算手段と、前記分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する他話題段落決定手段と、前記分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める文書ベクトル計算手段と、前記分類対象文書の文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する分類決定手段とを備えることを特徴とする分類装置。

【請求項4】　学習用文書と該学習用文書から選出された有効語を集めて作成した有効語辞書とを用いて、分類対象文書をユーザの意図に沿って分類する文書自動分類方法において、前記分類対象文書が属するカテゴリを決定するための基準を求めるための学習工程と、前記基準を用いて前記分類対象文書が属するカテゴリを決定するための分類工程とを有し、前記学習工程は、前記学習用文書について前記有効語辞書を参照して段落ベクトルを求める工程と、前記学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する工程と、前記学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める工程と、前記学習用文書の文書ベクトルを用いて前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求める工程とを含み、前記分類工程は、前記分類対象文書について前記有効語辞書を参照して段落ベクトルを求める工程と、前記分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する工程と、前記分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める工程と、前記分類対象文書の文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する工程とを含むことを特徴とする文書自動分類方法。

【請求項5】　分類対象文書をユーザの意図に沿って分類する文書自動分類システムに用いられる、前記分類対象文書が属するカテゴリを決定するための基準を求めるための学習方法において、学習用文書を保持する工程と、前記学習用文書から有効語を選出し、該選出された有効語を集めて有効語辞書を作成する工程と、前記学習用文書について前記有効語辞書を参照して段落ベクトルを求める工程と、前記学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する工程と、前記学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める工程と、前記学習用文書の文書ベクトルを用いて前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求める工程とを含むことを特徴とする学習方法。

【請求項6】　分類対象文書をユーザの意図に沿って分類する文書自動分類システムに請求項5記載の学習方法

とともに用いられる、前記分類対象文書が属するカテゴリを決定するための分類方法において、前記分類対象文書を保持する工程と、前記分類対象文書について前記有効語辞書を参照して段落ベクトルを求める工程と、前記分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する工程と、前記分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める工程と、前記分類対象文書の文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する工程とを含むことを特徴とする分類方法。

【請求項7】　学習用文書と該学習用文書から選出された有効語を集めて作成した有効語辞書とを用いて、分類対象文書をユーザの意図に沿って分類する文書自動分類システムを構築するためのプログラムを格納した記憶媒体において、前記プログラムは、前記分類対象文書が属するカテゴリを決定するための基準を求めるための学習プログラムと、前記基準を用いて前記分類対象文書が属するカテゴリを決定するための分類プログラムとを有し、前記学習プログラムは、前記学習用文書について前記有効語辞書を参照して段落ベクトルを求めるモジュールと、前記学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定するモジュールと、前記学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求めるモジュールと、前記学習用文書の文書ベクトルを用いて前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるモジュールとを含み、前記分類プログラムは、前記分類対象文書について前記有効語辞書を参照して段落ベクトルを求めるモジュールと、前記分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定するモジュールと、前記分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求めるモジュールと、前記分類対象文書の文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定するモジュールとを含むことを特徴とする記憶媒体。

【請求項8】　分類対象文書をユーザの意図に沿って分類する文書自動分類システムに用いられ、前記分類対象文書が属するカテゴリを決定するための基準を求める学習装置を構築するための学習プログラムを格納した記憶媒体において、前記学習プログラムは、学習用文書を保持するモジュールと、前記学習用文書から有効語を選出し、該選出された有効語を集めて有効語辞書を作成するモジュールと、前記学習用文書について前記有効語辞書を参照して段落ベクトルを求めるモジュールと、前記学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定するモジュールと、前記学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求めるモジュールと、前記学習用文書の文書ベクトルを用いて前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるモジュールとを含むことを特徴とする記憶媒体。

【請求項9】　分類対象文書をユーザの意図に沿って分類する文書自動分類システムに請求項8記載の記憶媒体とともに用いられる、前記分類対象文書が属するカテゴリを決定する分類装置を構築するための分類プログラムを格納した記憶媒体において、前記分類プログラムは、前記分類対象文書を保持するモジュールと、前記分類対象文書について前記有効語辞書を参照して段落ベクトルを求めるモジュールと、前記分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定するモジュールと、前記分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求めるモジュールと、前記分類対象文書の文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定するモジュールとを含むことを特徴とする記憶媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類装置、それに用いられる学習装置および分類装置と、文書自動分類方法、それに用いられる学習方法および分類方法と、文書自動分類装置を構築するための記憶媒体とに関する。

【0002】

【従来の技術】分類対象文書をユーザの意図に沿って分類する方法の一つとして、ベクトル空間モデルを利用した方法がある。このベクトル空間モデルでは、分類に有用な語や文書、カテゴリをベクトルで表現し、ベクトルの方向から文書が属するカテゴリを決定する。このベクトル空間モデルを利用した文書自動分類処理は、学習フェーズと分類フェーズとに分けられる。学習フェーズでは、予め正しく分類された学習用文書から分類に有用な語（以下、有効語という）を選出し、各有効語をベクトル表現して有効語辞書を作成する。また、学習用文書をベクトル表現して、各カテゴリの特徴を表すフォルダベクトルを算出する。分類フェーズでは、学習フェーズで得られた有効語辞書を用いて分類対象文書をベクトルで表現し（以下、文書ベクトルという）、この文書ベクトルとフォルダベクトルとを比較し、該比較結果に応じて

分類対象文書が属するカテゴリを決定する。

【0003】この方法を採用した文書自動分類装置の構成について図8ないし図11を参照しながら説明する。図8は従来の文書自動分類装置の構成を示すブロック図、図9は図8の文書自動分類装置における学習フェーズの処理手順を示すフローチャート、図10は図8の文書自動分類装置における分類フェーズの処理手順を示すフローチャート、図11は図8の文書自動分類装置における分類フェーズで求められた文書ベクトルの例を示す図である。

【0004】文書自動分類装置は、図8に示すように、学習用文書を保持する学習用文書保持部501と、分類対象文書を保持する分類対象文書保持部502と、学習用文書から有効語を選出し、この有効語を集めて有効語辞書を作成する有効語辞書作成部503と、有効語辞書を保持する有効語辞書保持部504と、学習用文書と分類対象文書とのそれぞれについて、有効語辞書を参照して文書ベクトルを求める文書ベクトル計算部505と、学習用文書と分類対象文書とのそれぞれについて求められた文書ベクトルを保持する文書ベクトル保持部506とを備える。

【0005】文書ベクトル保持部506に保持された学習用文書の文書ベクトルはフォルダベクトル計算部507に与えられ、フォルダベクトル計算部507は学習用文書の文書ベクトルを用いて各カテゴリのフォルダベクトルを求める。求められた各カテゴリのフォルダベクトルは、フォルダベクトル保持部508に保持される。

【0006】フォルダベクトル保持部508に保持された各カテゴリのフォルダベクトルは、文書ベクトル保持部506に保持された分類対象文書の文書ベクトルとともに分類決定部509に与えられ、分類決定部509は分類対象文書の文書ベクトルと各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定する。この決定された分類対象文書のカテゴリは分類結果保持部510に保持される。

【0007】次に、文書自動分類装置における学習フェーズの処理手順について図9を参照しながら説明する。

【0008】まず、ステップS601において学習要文書に含まれる語の中から、分類に有用な語を有効語として選定し、続くステップS602で、選定した有効語を出現頻度や他の有効語との共起状況などによりベクトル表現し、有効語辞書として保持する。

【0009】次いで、ステップS603に進み、学習用文書から有効語を抽出し、続くステップS604で、有効語辞書を参照して取り出した有効語のベクトルの平均を取り、このベクトルの平均から学習用文書の文書ベクトルを求める。そして、ステップS605で、学習用文書における各カテゴリに属する文書の文書ベクトルの平均を取り、該文書のベクトルの平均からフォルダベクトルを求め、本処理を終了する。

【0010】この学習フェーズが終了すると、分類フェーズが開始される。この分類フェーズの処理手順について図10を参照しながら説明する。

【0011】分類フェーズでは、まずステップS701において分類対象文書から有効語を取り出し、続くステップS702で有効語辞書を参照して取り出した有効語のベクトルの平均を取り、このベクトルの平均から分類対象文書の文書ベクトルを求める。

【0012】次いで、ステップS703に進み、分類対象文書の文書ベクトルと学習フェーズで求められたフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定し、本処理を終了する。

【0013】

【発明が解決しようとする課題】しかし、上述した従来の文書自動分類装置では、学習用文書または分類対象文書から有効語を取り出す際に、それぞれの文書中に出現する全ての有効語を取り出し、それぞれの文書についてその全ての有効語を用いて文書ベクトルを求めるから、文書中に主題と異なる他話題が挿入されている場合には、文書ベクトルが主題からそれた方向を示すことがある。例えば、図11に示すように、有効語のベクトルa～fを有する文書においては、文書ベクトルが他の話題に出現する有効語のベクトルe，fに引っ張られ、主題に出現する有効語のベクトルa～dの方向（主題の方向）から文書ベクトルがそれてしまい、適正に文書の分類を行なうことができない。

【0014】本発明の目的は、主題と異なる他話題が出現する文書に対してその分類を適正に行うことができる文書自動分類装置、文書自動分類方法および記憶媒体を提供することにある。

【0015】本発明の他の目的は、主題と異なる他話題が出現する文書に対してその分類を適正に行うことが可能な文書自動分類システムを実現することができる学習装置、分類装置、学習方法、分類方法および記憶媒体を提供することにある。

【0016】

【課題を解決するための手段】請求項1記載の発明は、学習用文書と該学習用文書から選出された有効語を集めて作成した有効語辞書とを用いて、分類対象文書をユーザの意図に沿って分類する文書自動分類装置において、前記学習用文書と前記分類対象文書とのそれぞれについて、前記有効語辞書を参照して段落ベクトルを求める段落ベクトル計算手段と、前記学習用文書と前記分類対象文書とのそれぞれについて、その段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する他話題段落決定手段と、前記学習用文書と前記分類対象文書とのそれぞれについて、その他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める文書ベクトル計算手段と、前記学習用文書について求

められた文書ベクトルを用いて各カテゴリのフォルダベクトルを求めるフォルダベクトル計算手段と、前記分類対象文書について求められた文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する分類決定手段とを備えることを特徴とする。

【００１７】請求項２記載の発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類システムに用いられる、前記分類対象文書が属するカテゴリを決定するための基準を求めるための学習装置において、学習用文書を保持する学習用文書保持手段と、前記学習用文書から有効語を選出し、該選出された有効語を集めて有効語辞書を作成する有効語辞書作成手段と、前記学習用文書について前記有効語辞書を参照して段落ベクトルを求める段落ベクトル計算手段と、前記学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する他話題段落決定手段と、前記学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める文書ベクトル計算手段と、前記学習用文書の文書ベクトルを用いて前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるフォルダベクトル計算手段とを備えることを特徴とする。

【００１８】請求項３記載の発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類システムに請求項２記載の学習装置とともに用いられる、前記分類対象文書が属するカテゴリを決定するための分類装置において、前記分類対象文書を保持する分類対象文書保持手段と、前記分類対象文書について前記有効語辞書を参照して段落ベクトルを求める段落ベクトル計算手段と、前記分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する他話題段落決定手段と、前記分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める文書ベクトル計算手段と、前記分類対象文書の文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する分類決定手段とを備えることを特徴とする。

【００１９】請求項４記載の発明は、学習用文書と該学習用文書から選出された有効語を集めて作成した有効語辞書とを用いて、分類対象文書をユーザの意図に沿って分類する文書自動分類方法において、前記分類対象文書が属するカテゴリを決定するための基準を求めるための学習工程と、前記基準を用いて前記分類対象文書が属するカテゴリを決定するための分類工程とを有し、前記学習工程は、前記学習用文書について前記有効語辞書を参照して段落ベクトルを求める工程と、前記学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する工程と、前記学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める工程と、前記学習用文書の文書ベクトルを用いて前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求める工程とを含み、前記分類工程は、前記分類対象文書について前記有効語辞書を参照して段落ベクトルを求める工程と、前記分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する工程と、前記分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める工程と、前記分類対象文書の文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する工程とを含むことを特徴とする。

【００２０】請求項５記載の発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類システムに用いられる、前記分類対象文書が属するカテゴリを決定するための基準を求めるための学習方法において、学習用文書を保持する工程と、前記学習用文書から有効語を選出し、該選出された有効語を集めて有効語辞書を作成する工程と、前記学習用文書について前記有効語辞書を参照して段落ベクトルを求める工程と、前記学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する工程と、前記学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める工程と、前記学習用文書の文書ベクトルを用いて前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求める工程とを含むことを特徴とする。

【００２１】請求項６記載の発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類システムに請求項５記載の学習方法とともに用いられる、前記分類対象文書が属するカテゴリを決定するための分類方法において、前記分類対象文書を保持する工程と、前記分類対象文書について前記有効語辞書を参照して段落ベクトルを求める工程と、前記分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する工程と、前記分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める工程と、前記分類対象文書の文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定する工程とを含むことを特徴とする。

【００２２】請求項７記載の発明は、学習用文書と該学習用文書から選出された有効語を集めて作成した有効語

辞書とを用いて、分類対象文書をユーザの意図に沿って分類する文書自動分類システムを構築するためのプログラムを格納した記憶媒体において、前記プログラムは、前記分類対象文書が属するカテゴリを決定するための基準を求めるための学習プログラムと、前記基準を用いて前記分類対象文書が属するカテゴリを決定するための分類プログラムとを有し、前記学習プログラムは、前記学習用文書について前記有効語辞書を参照して段落ベクトルを求めるモジュールと、前記学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定するモジュールと、前記学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求めるモジュールと、前記学習用文書の文書ベクトルを用いて前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるモジュールとを含み、前記分類プログラムは、前記分類対象文書について前記有効語辞書を参照して段落ベクトルを求めるモジュールと、前記分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定するモジュールと、前記分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求めるモジュールと、前記分類対象文書の文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定するモジュールとを含むことを特徴とする。

【００２３】請求項８記載の発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類システムに用いられ、前記分類対象文書が属するカテゴリを決定するための基準を求める学習装置を構築するための学習プログラムを格納した記憶媒体において、前記学習プログラムは、学習用文書を学習用文書保持手段に保持するモジュールと、前記学習用文書から有効語を選出し、該選出された有効語を集めて有効語辞書を作成するモジュールと、前記学習用文書について前記有効語辞書を参照して段落ベクトルを求めるモジュールと、前記学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定するモジュールと、前記学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求めるモジュールと、前記学習用文書の文書ベクトルを用いて前記分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるモジュールとを含むことを特徴とする。

【００２４】請求項９記載の発明は、分類対象文書をユーザの意図に沿って分類する文書自動分類システムに請求項８記載の記憶媒体とともに用いられる、前記分類対象文書が属するカテゴリを決定する分類装置を構築するための分類プログラムを格納した記憶媒体において、前記分類プログラムは、前記分類対象文書を保持するモジュールと、前記分類対象文書について前記有効語辞書を参照して段落ベクトルを求めるモジュールと、前記分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定するモジュールと、前記分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求めるモジュールと、前記分類対象文書の文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定するモジュールとを含むことを特徴とする。

【００２５】

【発明の実施の形態】以下に本発明の実施の形態について図を参照しながら説明する。

【００２６】図１は本発明の文書自動分類装置の実施の一形態の機能構成を示すブロック図、図２は図１の文書自動分類装置のハードウェア構成を示すブロック図である。

【００２７】文書自動分類装置は、図１に示すように、学習用文書を保持する学習用文書保持部１０１と、分類対象文書を保持する分類対象文書保持部１０２と、学習用文書から選出された有効語を集めて有効語辞書を作成する有効語辞書作成部１０３と、有効語辞書を保持する有効語辞書保持部１０４と、学習用文書と分類対象文書とのそれぞれについて、有効語辞書を参照して段落ベクトルを求める段落ベクトル計算部１０５と、学習用文書と分類対象文書とのそれぞれについて求められた段落ベクトルを保持する段落ベクトル保持部１０６とを備える。

【００２８】段落ベクトル保持部１０６に保持された学習用文書と分類対象文書とのそれぞれの段落ベクトルは他話題段落決定部１０７に与えられ、他話題段落決定部１０７は、学習用文書と分類対象文書とのそれぞれについて、その段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する。この学習用文書と分類対象文書とのそれぞれについて決定された他話題段落は、他話題段落保持部１０８に保持される。

【００２９】他話題段落保持部１０８に保持された学習用文書と分類対象文書とのそれぞれの他話題段落は、段落ベクトル保持部１０６に保持された学習用文書と分類対象文書とのそれぞれの段落ベクトルとともに文書ベクトル計算部１０９に与えられる。文書ベクトル計算部１０９は、学習用文書と分類対象文書とのそれぞれについて、その他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める。学習用文書と分類対象文書とのそれぞれについて求められた文書ベクト

ルは文書ベクトル保持部１１０に保持される。

【００３０】文書ベクトル保持部１１０に保持された学習用文書の文書ベクトルはフォルダベクトル計算部１１１に与えられる。フォルダベクトル計算部１１１は学習用文書の文書ベクトルを用いて各カテゴリのフォルダベクトルを求め、求められた各カテゴリのフォルダベクトルはフォルダベクトル保持部１１２に保持される。

【００３１】フォルダベクトル保持部１１２に保持された各カテゴリのフォルダベクトルは、文書ベクトル保持部１１０に保持された分類対象文書の文書ベクトルとともに分類決定部１１３に与えられる。分類決定部１１３は、分類対象文書の文書ベクトルと各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定し、この決定された分類対象文書のカテゴリは分類結果保持部１１４に保持される。

【００３２】文書自動分類装置のハードウェア構成においては、図２に示すように、ＲＯＭ２０１に格納されている制御プログラムを実行して後述する制御（図３および図４に示す制御）を行う中央処理装置２０３が設けられている。中央処理装置２０３の演算処理の作業領域としてはＲＡＭ２０２が用いられ、また、ＲＡＭ２０２は、段落ベクトル保持部１０６、他話題段落保持部１０８、文書ベクトル保持部１１０、分類結果保持部１１４のための記憶領域を提供する。

【００３３】中央処理装置２０３には、ＲＯＭ２０１およびＲＡＭ２０２とともに、ハードディスク装置２０４がバス２０５を介して接続され、ハードディスク装置２０４は、学習用文書保持部１０１、分類対象文書保持部１０２、有効語辞書保持部１０４およびフォルダベクトル保持部１１２を構成する。なお、ハードディスク装置２０４に代えて、他の記憶媒体を用いて、学習用文書保持部１０１、分類対象文書保持部１０２、有効語辞書保持部１０４およびフォルダベクトル保持部１１２を構成することも可能である。

【００３４】次に、本文書自動分類装置が実行する処理について図３および図４を参照しながら説明する。図３は図１の文書自動分類装置における学習フェーズの処理手順を示すフローチャート、図４は図１の文書自動分類装置における分類フェーズの処理手順を示すフローチャート、図５は図１の文書自動分類装置における文書ベクトルの例を示す図である。

【００３５】本文書自動分類装置における処理は学習フェーズと分類フェーズとに分けられ、最初に、学習フェーズの処理手順について図３を参照しながら説明する。

【００３６】学習フェーズでは、図３に示すように、まずステップＳ３０１において学習用文書に含まれる語の中から、分類に有用な有効語として選定し、続くステップＳ３０２で、選定した有効語を出現頻度や他の有効語との共起状況などによりベクトル表現し、有効語辞書として保持する。

【００３７】次いで、ステップＳ３０３に進み、学習用文書から有効語を抽出し、続くステップＳ３０４で、有効語辞書を参照して取り出した有効語のベクトルの平均を段落毎に算出し、各段落の段落ベクトルとする。

【００３８】次いで、ステップＳ３０５に進み、各文書毎に段落ベクトルの分布を調べ、他の段落ベクトルと極端に違う方向の段落ベクトルがあれば、その段落ベクトルの段落を主題とは異なる話題を表す他話題段落として決定する。また、各段落ベクトルにおいてそれぞれの向きが近接しているときには、全ての段落ベクトルが主題を表していると判断する。すなわち全ての段落を他話題段落ではないとする。例えば、各段落ベクトル毎に他の段落ベクトルとの余弦値の総和を求め、この総和が他より極端に小さい段落ベクトルがあれば、該段落ベクトルの段落を他話題段落として決定する。ここで、他の段落ベクトルとの余弦値の総和が正規分布に従うとすれば、該分布の数％以下を示す段落ベクトルの段落を他話題段落として定義することが可能である。

【００３９】他話題段落を決定すると、ステップＳ３０６に進み、他話題段落の段落ベクトルを段落ベクトル群から除去し、残りの段落ベクトルの平均を取って文書ベクトルとする。続くステップＳ３０７では、学習用文書における各カテゴリに属する文書の文書ベクトルの平均を取り、該文書ベクトルの平均からフォルダベクトルを求め、本処理を終了する。

【００４０】この学習フェーズが終了すると、分類フェーズが開始される。この分類フェーズの処理手順について図４を参照しながら説明する。

【００４１】分類フェーズでは、図４に示すように、まずステップＳ４０１において分類対象文書から有効語を取り出し、続くステップＳ４０２で有効語辞書を参照して取り出した有効語のベクトルの平均を段落毎に取り、各段落のベクトルの平均から分類対象文書の文書ベクトルを求める。

【００４２】次いで、ステップＳ４０３に進み、各文書毎に段落ベクトルの分布を調べて分類対象文書の他話題段落を決定する。この他話題段落の決定処理は上述の学習フェーズにおけるステップＳ３０５の処理内容と同じであり、その説明は省略する。分類対象文書の文書ベクトルと学習フェーズで求められたフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定し、本処理を終了する。

【００４３】例えば、図１１に示すように、主題と異なる他話題が出現する文書において、有効語のベクトルａ，ｂが段落Ａに、有効語のベクトルｃ，ｄが段落Ｂに、有効語のベクトルｅ，ｆが段落Ｃにそれぞれ出現したとすると、各段落Ａ，Ｂ，Ｃの段落ベクトルの内の段落Ｃが、図５に示すように、主題と異なる他話題を表す他話題段落として決定され、この段落Ｃの段落ベクトルを除去して各段落Ａ，Ｂの段落ベクトルから文書ベクト

ルが求められる。よって、求められた文書ベクトルが主題の方向にほぼ一致することになり、適正に文書の分類を行うことができる。

【００４４】以上より、本実施の形態では、各段落の段落ベクトルを求めてその分布からはずれている段落を他話題段落として除去し、残りの段落から文書ベクトルを求めることにより、主題の方向をほぼ示すような文書ベクトルが得られ、主題と異なる他話題が出現する文書に対してその分類を適正に行うことができる。

【００４５】なお、本実施の形態では、他話題段落を段落ベクトル間の余弦値の総和に基づき決定する例を示したが、これに限定されるものではなく、他の値を用いて他話題段落を決定することも可能である。また、段落ベクトルの分布において、該分布からはずれているか否かの決定に正規分布に従う値を基準値として決定する方法を示したが、これに限定されるものではない。

【００４６】また、本実施の形態では、学習フェーズにおいて、段落ベクトルの計算、他話題段落の決定、文書ベクトルの計算の各処理を学習用文書に含まれる全ての文書に対して実行し、その後に次の処理を実行するように設定しているが、これに限定されるものではなく、１文書づつ各処理を実行するように設定することも可能である。

【００４７】さらに、本実施の形態では、段落単位で話題を取り扱っているが、これに限定されるものではなく、文や節など、他の文章単位で扱うことも可能である。

【００４８】さらに、本実施の形態では、文書ベクトルを平均してフォルダベクトルを求めるように説明しているが、各カテゴリの段落ベクトルの平均をフォルダベクトルとしてもよい。この場合、学習フェーズにおいて、文書ベクトルを求める必要はない。

【００４９】さらに、本実施の形態では、上述の処理（各ブロックの機能）を実行するためのプログラムをＲＯＭに格納した例を示したが、他の記憶媒体を用いて上記プログラムを供給するように構成することも可能である。また、各ブロックの機能をそれぞれ有する回路構成により本装置を構成することも可能である。

【００５０】さらに、本装置をコンピュータなどの情報処理装置上に構築することも可能である。この場合、上述の処理（各ブロックの機能）を実行するためのプログラムを格納した記憶媒体を準備し、ＣＰＵなどが該記憶媒体から上記プログラムを読み出して実行することにより、文書自動分類装置が構成される。上記プログラムを供給するための記憶媒体としては、フロッピーディスク、ハードディスク、光ディスク、光磁気ディスク、ＣＤＲＯＭ、ＣＤ－Ｒ、磁気テープ、不揮発性メモリカード、ＲＯＭなどを用いることができる。なお、上記プログラムの実行により文書自動分類装置を構成する場合には、コンピュータ上で稼働しているＯＳが上記プログラムに含まれる処理の一部または全てを実行するように構成されている場合も含まれる。また、記憶媒体から供給されたプログラムがコンピュータに搭載された拡張機能ボードまたは接続された周辺拡張ユニットに書き込まれた後に、拡張機能ボードまたは周辺拡張ユニットに設けられたＣＰＵが書き込まれたプログラムを実行する場合も含まれる。

【００５１】さらに、本発明の原理は、複数の機器からなるシステム、ひとつの機器からなる装置のいずれにも適用することが可能である。

【００５２】さらに、本実施の形態では、学習フェーズと分類フェーズとを一つの装置上で行う例を説明したが、これに限定されるものではなく、例えば、学習フェーズを行う装置と、分類フェーズを行う装置とを準備し、それぞれの装置を用いて文書の分類を行うように構成することもできる。この場合、学習フェーズを行う装置により、有効語辞書を作成しまたフォルダベクトルを求め、この有効語辞書およびフォルダベクトルを可搬記憶媒体または通信により、分類フェーズを行う装置に供給して分類を行う方法が用いられる。

【００５３】この学習フェーズを行う装置および分類フェーズを行う装置について図６および図７を参照しながら説明する。図６は本発明の学習装置の実施の一形態の構成を示すブロック図、図７は本発明の分類装置の実施の一形態の構成を示すブロック図である。

【００５４】学習フェーズを行う装置は、図６に示すように、学習用文書を保持する学習用文書保持部１００１と、学習用文書から選出された有効語を集めて有効語辞書を作成する有効語辞書作成部１００２と、有効語辞書を保持する有効語辞書保持部１００３と、学習用文書について、有効語辞書を参照して段落ベクトルを求める段落ベクトル計算部１００４と、学習用文書について求められた段落ベクトルを保持する段落ベクトル保持部１００５とを備える。

【００５５】段落ベクトル保持部１００５に保持された学習用文書の段落ベクトルは他話題段落決定部１００６に与えられ、他話題段落決定部１００６は、学習用文書について、その段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する。この学習用文書と分類対象文書とのそれぞれについて決定された他話題段落は、他話題段落保持部１００７に保持される。

【００５６】他話題段落保持部１００７に保持された学習用文書の他話題段落は、段落ベクトル保持部１００５に保持された学習用文書の段落ベクトルとともに文書ベクトル計算部１００８に与えられる。文書ベクトル計算部１００８は、学習用文書について、その他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める。学習用文書について求められた文書ベクトルは文書ベクトル保持部１００９に保持される。

【００５７】文書ベクトル保持部１００９に保持された学習用文書の文書ベクトルはフォルダベクトル計算部１０１０に与えられる。フォルダベクトル計算部１０１０は学習用文書の文書ベクトルを用いて各カテゴリのフォルダベクトルを求め、求められた各カテゴリのフォルダベクトルはフォルダベクトル保持部１０１１に保持される。

【００５８】フォルダベクトル保持部１０１１に保持され各カテゴリのフォルダベクトル、および有効語辞書保持部１００３に保持された有効語辞書は、可搬記憶媒体に記憶されて分類フェーズを行う装置に供給され、または通信により分類フェーズを行う装置に供給される。

【００５９】分類フェーズを行う装置は、図７に示すように、分類対象文書を保持する分類対象文書保持部１１０１と、学習フェーズを行う装置から可搬記憶媒体または通信を介して供給された有効語辞書を保持する有効語辞書保持部１１０２と、学習フェーズを行う装置から可搬記憶媒体または通信を介して供給されたフォルダベクトルを保持するフォルダベクトル保持部１１０９と、分類対象文書について、有効語辞書を参照して段落ベクトルを求める段落ベクトル計算部１１０３と、分類対象文書について求められた段落ベクトルを保持する段落ベクトル保持部１１０４とを備える。

【００６０】段落ベクトル保持部１１０４に保持された分類対象文書の段落ベクトルは他話題段落決定部１１０５に与えられ、他話題段落決定部１１０５は、分類対象文書について、その段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する。この分類対象文書について決定された他話題段落は、他話題段落保持部１１０６に保持される。

【００６１】他話題段落保持部１１０６に保持された分類対象文書の他話題段落は、段落ベクトル保持部１１０４に保持された分類対象文書の段落ベクトルとともに文書ベクトル計算部１１０７に与えられる。文書ベクトル計算部１１０７は、分類対象文書について、その他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める。分類対象文書について求められた文書ベクトルは文書ベクトル保持部１１０８に保持される。

【００６２】文書ベクトル保持部１１０８に保持された分類対象文書の文書ベクトルは、フォルダベクトル保持部１１０９に保持された各カテゴリのフォルダベクトルとともに分類決定部１１１０に与えられる。分類決定部１１１０は、分類対象文書の文書ベクトルと各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定し、この決定された分類対象文書のカテゴリは分類結果保持部１１１１に保持される。

【００６３】

【発明の効果】以上に説明したように、請求項１記載の文書自動分類装置によれば、学習用文書と分類対象文書とのそれぞれについて、有効語辞書を参照して段落ベクトルを求める段落ベクトル計算手段と、学習用文書と分類対象文書とのそれぞれについて、その段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する他話題段落決定手段と、学習用文書と分類対象文書とのそれぞれについて、その他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める文書ベクトル計算手段と、学習用文書について求められた文書ベクトルを用いて各カテゴリのフォルダベクトルを求めるフォルダベクトル計算手段と、分類対象文書について求められた文書ベクトルと各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定する分類決定手段とを備えるから、主題の方向をほぼ示すような分類対象文書の文書ベクトルと、主題の方向を適正に示す各カテゴリのフォルダベクトルとが得られ、この文書ベクトルと各カテゴリのフォルダベクトルとの比較により、主題と異なる他話題が出現する文書に対してその分類を適正に行うことができる。

【００６４】請求項２記載の学習装置によれば、学習用文書を保持する学習用文書保持手段と、学習用文書から有効語を選出し、該選出された有効語を集めて有効語辞書を作成する有効語辞書作成手段と、学習用文書について有効語辞書を参照して段落ベクトルを求める段落ベクトル計算手段と、学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する他話題段落決定手段と、学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める文書ベクトル計算手段と、学習用文書の文書ベクトルを用いて分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるフォルダベクトル計算手段とを備えるから、主題の方向を適正に示す各カテゴリのフォルダベクトルが得られ、主題と異なる他話題が出現する文書に対してその分類を適正に行うことが可能な文書自動分類システムを実現することができる。

【００６５】請求項３記載の分類装置によれば、分類対象文書を保持する分類対象文書保持手段と、分類対象文書について有効語辞書を参照して段落ベクトルを求める段落ベクトル計算手段と、分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する他話題段落決定手段と、分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める文書ベクトル計算手段と、分類対象文書の文書ベクトルと各カテゴリのフォルダベクトルとを

比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定する分類決定手段とを備えるから、主題の方向をほぼ示すよう分類対象文書の文書ベクトルが得られ、この文書ベクトルと主題の方向を適正に示す各カテゴリのフォルダベクトルとの比較により、主題と異なる他話題が出現する文書に対してその分類を適正に行うことが可能な文書自動分類システムを実現することができる。

【００６６】請求項４記載の文書自動分類方法によれば、分類対象文書が属するカテゴリを決定するための基準を求めるための学習工程と、基準を用いて分類対象文書が属するカテゴリを決定するための分類工程とを有し、学習工程が、学習用文書について有効語辞書を参照して段落ベクトルを求める工程と、学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する工程と、学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める工程と、学習用文書の文書ベクトルを用いて分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求める工程とを含み、分類工程が、分類対象文書について有効語辞書を参照して段落ベクトルを求める工程と、分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する工程と、分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める工程と、分類対象文書の文書ベクトルと各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定する工程とを含むから、主題の方向をほぼ示すような分類対象文書の文書ベクトルと、主題の方向を適正に示す各カテゴリのフォルダベクトルとが得られ、この文書ベクトルと各カテゴリのフォルダベクトルとの比較により、主題と異なる他話題が出現する文書に対してその分類を適正に行うことができる。

【００６７】請求項５記載の学習方法によれば、学習用文書を学習用文書保持手段に保持する工程と、学習用文書から有効語を選出し、該選出された有効語を集めて有効語辞書を作成する工程と、学習用文書について有効語辞書を参照して段落ベクトルを求める工程と、学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する工程と、学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める工程と、学習用文書の文書ベクトルを用いて分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求める工程とを含むから、主題の方向を適正に示す各カテゴリのフォルダベクトルが得られ、主題と異なる他話題が出現する文書に対してその分類を適正に行うことが可能な文書自動分類システムを実現することができる。

【００６８】請求項６記載の分類方法によれば、分類対象文書を保持する工程と、分類対象文書について有効語辞書を参照して段落ベクトルを求める工程と、分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定する工程と、分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求める工程と、分類対象文書の文書ベクトルと各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて分類対象文書が属するカテゴリを決定する工程とを含むから、主題の方向をほぼ示すよう分類対象文書の文書ベクトルが得られ、この文書ベクトルと主題の方向を適正に示す各カテゴリのフォルダベクトルとの比較により、主題と異なる他話題が出現する文書に対してその分類を適正に行うことが可能な文書自動分類システムを実現することができる。

【００６９】請求項７記載の記憶媒体によれば、前記プログラムが、分類対象文書が属するカテゴリを決定するための基準を求めるための学習プログラムと、基準を用いて分類対象文書が属するカテゴリを決定するための分類プログラムとを有し、学習プログラムが、学習用文書について有効語辞書を参照して段落ベクトルを求めるモジュールと、学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定するモジュールと、学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求めるモジュールと、学習用文書の文書ベクトルを用いて分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるモジュールとを含み、分類プログラムが、分類対象文書について有効語辞書を参照して段落ベクトルを求めるモジュールと、分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定するモジュールと、分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求めるモジュールと、分類対象文書の文書ベクトルと各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定するモジュールとを含むから、主題の方向をほぼ示すような分類対象文書の文書ベクトルと、主題の方向を適正に示す各カテゴリのフォルダベクトルとが得られ、この文書ベクトルと各カテゴリのフォルダベクトルとの比較により、主題と異なる他話題が出現する文書に対してその分類を適正に行うことができる。

【００７０】請求項８記載の記憶媒体によれば、学習プログラムが、学習用文書を保持するモジュールと、学習用文書から有効語を選出し、該選出された有効語を集め

て有効語辞書を作成するモジュールと、学習用文書について有効語辞書を参照して段落ベクトルを求めるモジュールと、学習用文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定するモジュールと、学習用文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求めるモジュールと、学習用文書の文書ベクトルを用いて分類対象文書が属するカテゴリを決定するための基準となる各カテゴリのフォルダベクトルを求めるモジュールとを含むから、主題の方向を適正に示す各カテゴリのフォルダベクトルが得られ、主題と異なる他話題が出現する文書に対してその分類を適正に行うことが可能な文書自動分類システムを実現することができる。

【0071】請求項9記載の記憶媒体によれば、分類プログラムが、分類対象文書を保持するモジュールと、分類対象文書について有効語辞書を参照して段落ベクトルを求めるモジュールと、分類対象文書の段落ベクトルの分布から主題とは異なる話題を表す他話題段落を決定するモジュールと、分類対象文書の他話題段落を参照してその段落ベクトルの中から有効な段落ベクトルを取り出し、該取り出した段落ベクトルを用いて文書ベクトルを求めるモジュールと、分類対象文書の文書ベクトルと前記各カテゴリのフォルダベクトルとを比較し、該比較結果に応じて前記分類対象文書が属するカテゴリを決定するモジュールとを含むから、主題の方向をほぼ示すよう分類対象文書の文書ベクトルが得られ、この文書ベクトルと主題の方向を適正に示す各カテゴリのフォルダベクトルとの比較により、主題と異なる他話題が出現する文書に対してその分類を適正に行うことが可能な文書自動分類システムを実現することができる。

【図面の簡単な説明】

【図1】本発明の文書自動分類装置の実施の一形態の機能構成を示すブロック図である。

【図2】図1の文書自動分類装置のハードウェア構成を示すブロック図である。

【図3】図1の文書自動分類装置における学習フェーズの処理手順を示すフローチャートである。

【図4】図1の文書自動分類装置における分類フェーズの処理手順を示すフローチャートである。

【図5】図1の文書自動分類装置における文書ベクトルの例を示す図である。

【図6】本発明の学習装置の実施の一形態の構成を示すブロック図である。

【図7】本発明の分類装置の実施の一形態の構成を示すブロック図である。

【図8】従来の文書自動分類装置の構成を示すブロック図である。

【図9】図8の文書自動分類装置における学習フェーズの処理手順を示すフローチャートである。

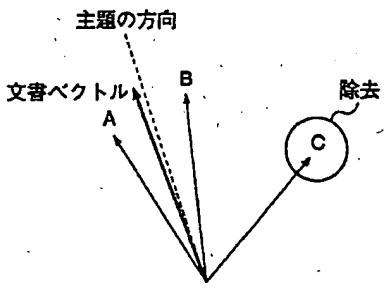
【図10】図8の文書自動分類装置における分類フェーズの処理手順を示すフローチャートである。

【図11】図8の文書自動分類装置における分類フェーズで求められた文書ベクトルの例を示す図である。

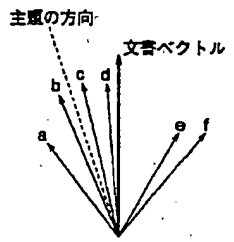
【符号の説明】

101，1001　学習用文書保持部
102，1101　分類対象文書保持部
103，1002　有効語辞書作成部
104，1003，1102　有効語辞書保持部
105，1004，1103　段落ベクトル計算部
106，1005，1104　段落ベクトル保持部
107，1006，1105　他話題段落決定部
108，1007，1106　他話題段落保持部
109，1008，1107　文書ベクトル計算部
110，1009，1108　文書ベクトル保持部
111，1010，　フォルダベクトル計算部
112，1011，1109　フォルダベクトル保持部
113、1110　分類決定部
114、1111　分類結果保持部
201　ROM
202　RAM
203　中央処理装置
204　ハードディスク装置

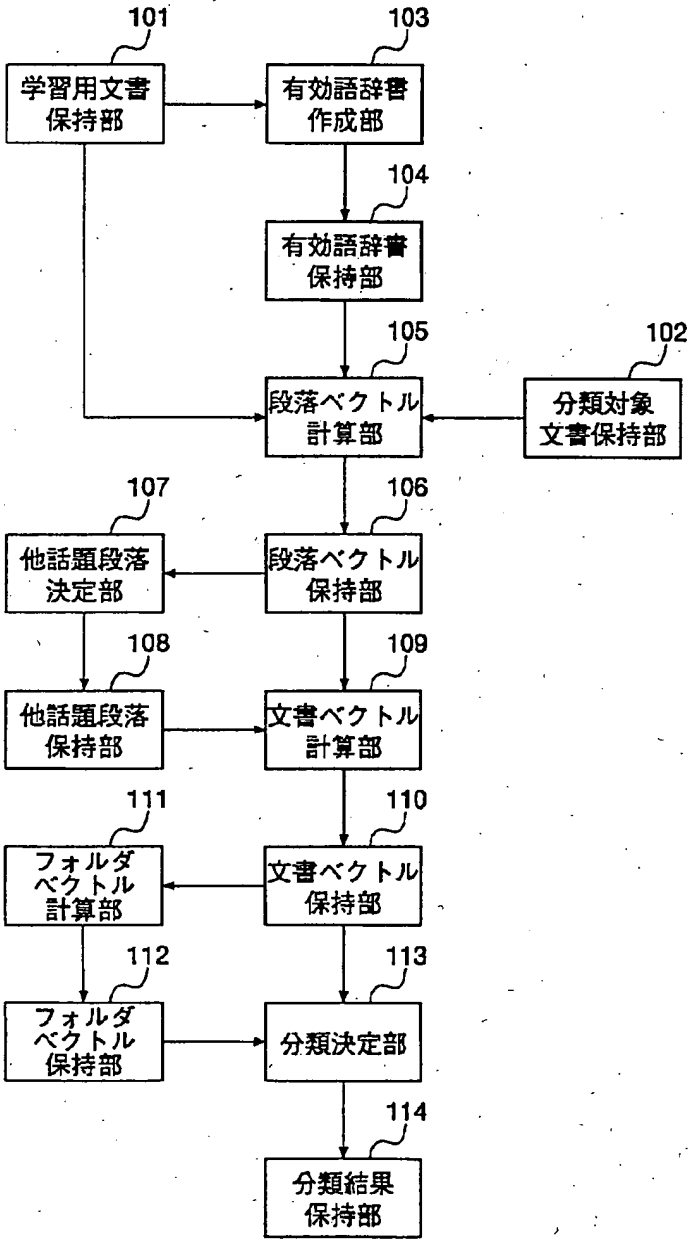
【図5】



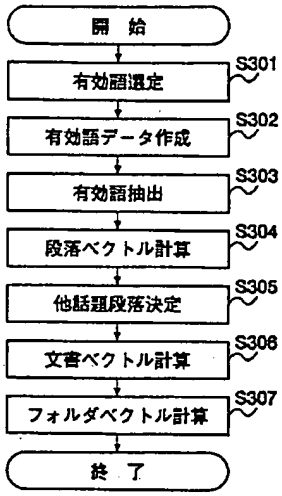
【図11】



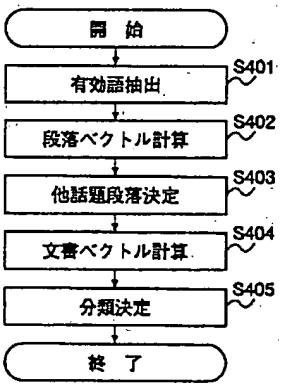
【図1】


101
学習用文書保持部
103
有効語辞書作成部
104
有効語辞書保持部
105
段落ベクトル計算部
102
分類対象文書保持部
107
他話題段落決定部
106
段落ベクトル保持部
108
他話題段落保持部
109
文書ベクトル計算部
111
フォルダベクトル計算部
110
文書ベクトル保持部
112
フォルダベクトル保持部
113
分類決定部
114
分類結果保持部

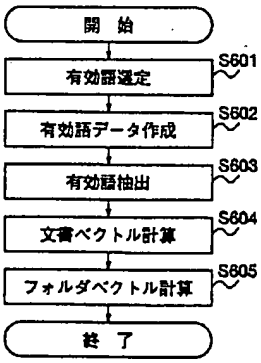
【図3】


開　始
有効語選定
S301
有効語データ作成
S302
有効語抽出
S303
段落ベクトル計算
S304
他話題段落決定
S305
文書ベクトル計算
S306
フォルダベクトル計算
S307
終　了

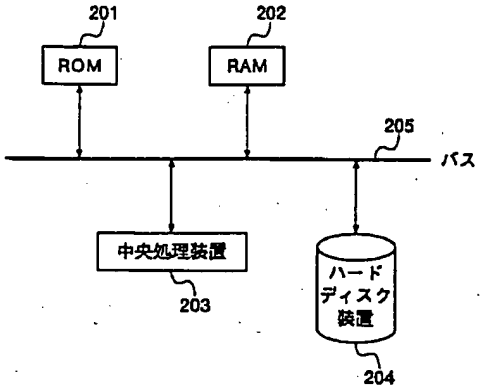
【図4】


開　始
有効語抽出
S401
段落ベクトル計算
S402
他話題段落決定
S403
文書ベクトル計算
S404
分類決定
S405
終　了

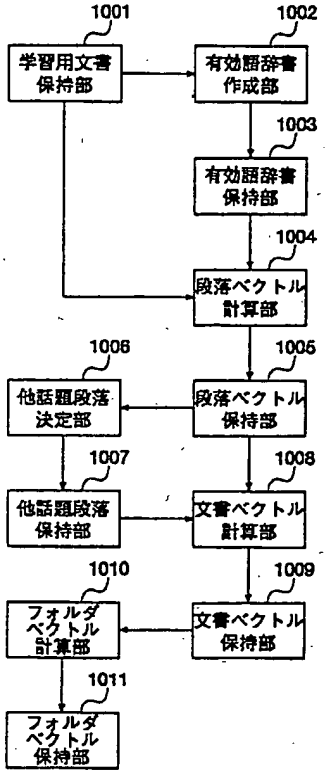
【図9】


開　始
有効語選定
S601
有効語データ作成
S602
有効語抽出
S603
文書ベクトル計算
S604
フォルダベクトル計算
S605
終　了

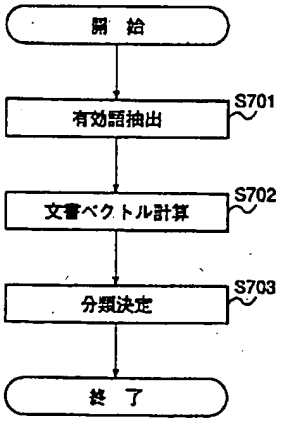
【図2】


201
ROM
202
RAM
205
バス
中央処理装置
203
ハードディスク装置
204

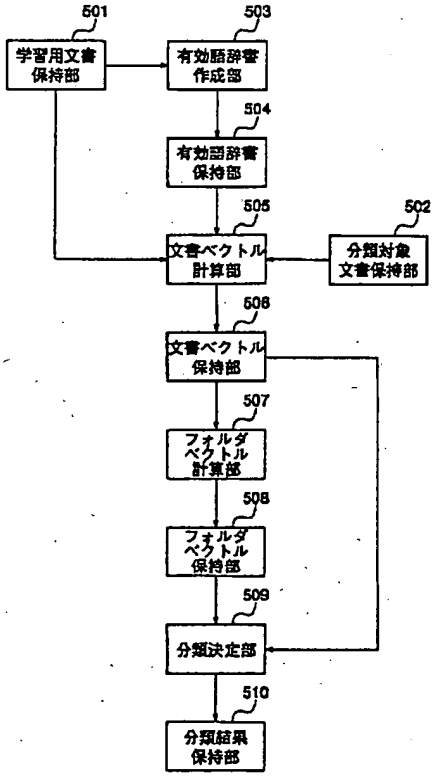
【図6】


1001
学習用文書保持部
1002
有効語辞書作成部
1003
有効語辞書保持部
1004
段落ベクトル計算部
1005
段落ベクトル保持部
1006
他話題段落決定部
1007
他話題段落保持部
1008
文書ベクトル計算部
1009
文書ベクトル保持部
1010
フォルダベクトル計算部
1011
フォルダベクトル保持部

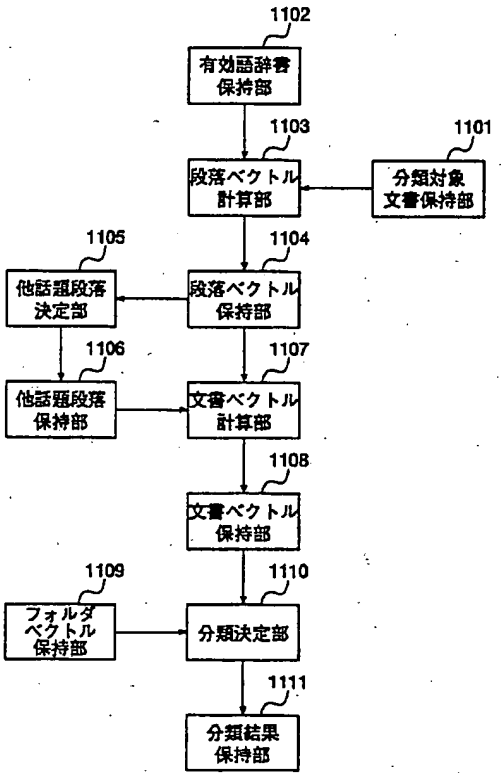
【図10】


開　始
有効語抽出
S701
文書ベクトル計算
S702
分類決定
S703
終　了

【図8】


501
学習用文書保持部
503
有効語辞書作成部
504
有効語辞書保持部
505
文書ベクトル計算部
502
分類対象文書保持部
506
文書ベクトル保持部
507
フォルダベクトル計算部
508
フォルダベクトル保持部
509
分類決定部
510
分類結果保持部

【図7】



フロントページの続き

(72)発明者　上田　隆也
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内

(72)発明者　池田　裕治
東京都大田区下丸子３丁目30番２号　キヤノン株式会社内